University of Málaga

Health Engineering

Laboratory Task

Classification

Author

Alejandro Dominguez Recio

Course

Intelligent Systems

Teachers

Enrique Domínguez Merino

Jesús de Benito Picazo

Introduction

In this practice we going to evaluate three diferents dataset with the classifier Naive Bayes. To evaluate which dataset is more optimal with this classifier we going to describe the performance measures seen in class like confusion matrix, acurracy, precision, fallout, recall, F-measure and area under ROC curve. Also we going to detail the characteristics of the different datasets.

How are we going to do it?

To do the different studies we will have the following structure in each of then.

➔ **Dataset context**

In this part we going to write a little introduction of the dataset context commenting on data type and prediction target.

➔ **Number of classes**

We going to describe the number of classes as if it is binary o multiclass. To do that we going to observe the atributtes weka panel and we going to select the class attribute and depending on the number of values that it takes, we will determine if it is binary or multiclass.

➔ **Number of attributes**

For the number of attributes we going to inspect the weka attributes panel or open the file in text mode and inspect the characteristics.

➔ **Number of samples or instances**

To know the number of samples we can proceed as in the previous section by inspecting the weka panels or opening the document in text mode and seeing its characteristics.

➔ Values of permance measures

Accuracy

We can obtain the acurracy in two ways, one of them is directly from the classifier out and the another one is by calculating the number of correctly predicted examples divided by the total number of examples.

Precision

This measurement shows the positive predictive value, higher is bettter. We are going to obtain the necessary values from the confusion matrix and perform the following calculation TP/ (TP+FP).

Fallout

This measurement shows the false positive rate, lower is better. We are going to obtain the necessary values from the confusion matrix and perform the following calculation FP/(FP+TN).

Recall

This measurement shows the true positive rate, higher is better. We are going to obtain the necessary values from the confusion matrix and perform the following calculation TP/(TP+FN).

F-measure

This measurement provides a single score that balances both the concers of precision and recall in one number, higher is better. We are going to obtain the necessary values of the previous measurements, precision and recall, and perform the following calculation 2*(Precision*Recall)/(Precision+Recall).

➔ **ROC curve and the area under the curve**

We are going to obtain this measurement by visualizing the threshold curve in Weka. The area under the ROC curve is a number in the interval [0,1], a higher value is better. This measure show the trade-off between the ratios of false positives and false negatives. This measurement is very useful when we faced with unbalanced data.

*The previous points will be repeated for each dataset

➔ **Evaluation of the results**

In this section we are goint to compare the results of the different performance measures.

➔ **Differences between datasets**

Here we are going to describe the main differences in the data of the different datasets.

➔ **Last conclusions**

Finally, we are going explain possible reasons why some datasets perform better with the Naive Bayes classifier than others. We will support our conclusions on the performance measures taken and on the characteristics of the datasets

BREAST CANCER DATASET

➔ Introduction

The main idea of this classification problem is given a dataset about patients with a series of characteristics which determine if they have recurrence events or no recurrence events of breast cancer train a Naive Bayes classification model and evaluate its performance.

In this dataset we have 286 instances in total of which 201 of one class and 85 instances of another class. Each instance has ten attributes, one of them is the class attribute. Because of we have only two classes we are faced with a binary classification problem.

➔ Number of classes

In this case our class attribute can take only **two *values***, *recurrence events* and *no recurrence events*. Because of that we are faced to a ***binary classification problem***.

➔ Number of attributes

In this case we have **ten attributes** which are *class, age, menopause, tumor-size, inv-nodes, node-caps, deg-malig, breast, breast-quad* and *irradiat.*

➔ Number of samples

In this case in particular the number of samples or instances is **286.**

➔ **Values of permance measures**

Confusion Matrix

=== Confusion Matrix ===

a b ← classified as

171 30 | a = no-recurrence-events

48 37 | b = recurrence-events

Accuracy

Correctly Classified Instances 208

Incorrectly Classified Instances 78

Total instances 286

Acurracy = 208/286 = **72.7273 %**

Precision

TP/(TP+FP) = 171/(171+30) = **0,8507 %**

Fallout

FP/(FP+TN) = 30/(30+37) = **0,4477 %**

Recall

TP/(TP+FN) = 171/(171+48) = **0,7808 %**

F-measure

2*(Precision*Recall)/(Precision+Recall) = 2*(0,8507*0,7808)/(0,8507+0,7808) = **1,6315**

➔ ROC curve and the area under the curve

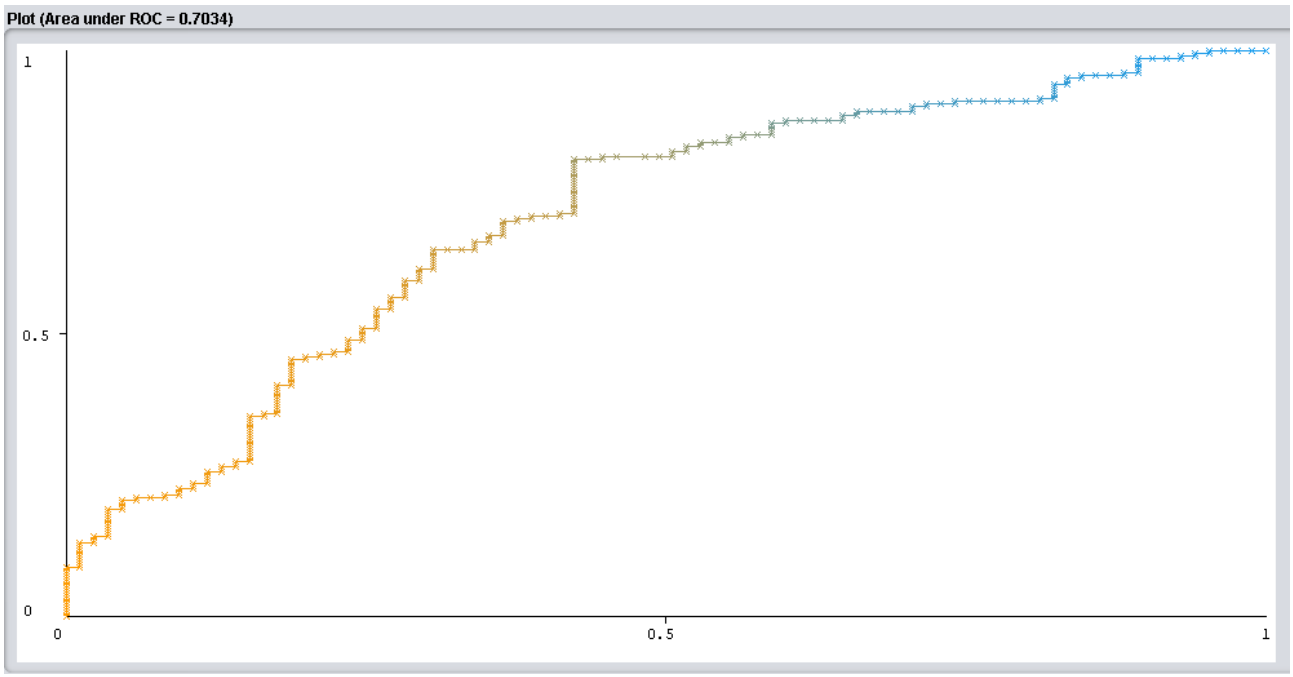

Ilustración 1: Graphic ROC Breast Cancer

HEPATITIS DATASET

➔ Introduction

The main idea of this classification problem is given a dataset about patients which suffer from hepatitis and they a series of characteristics which determine if they died or live because train a Naive Bayes classification model and evaluate its performance.

In this dataset we have 155 instances in total of which 129 of one class and 26 instances of another class. Each instance has 20 attributes, one of them is the class attribute. Because of we have only two classes we are faced with a binary classification problem.

➔ Number of classes

In this case our class attribute can take only ***two values***, *die* and *live*. Because of that we are faced to a ***binary classification problem***.

➔ Number of attributes

In this case we have **twenty attributes** which are *age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver big, liver firm, spleen palpable, spiders, ascites, varices, bilirubin, alk phosphate, sgot, albumin, protime, histology* and *class.*

➔ Number of samples

In this case in particular the number of samples or instances is **155.**

➔ **Values of permance measures**

Confusion Matrix

=== Confusion Matrix ===

a	b	← classified as
22	10	a = die
16	107	b = live

Accuracy

Correctly Classified Instances 129

Incorrectly Classified Instances 26

Total instances 155

Acurracy = 129/155 = **83,2258 %**

Precision

TP/(TP+FP) = 22/(22+10) = **0,6875 %**

Fallout

FP/(FP+TN) = 10/(10+107) = **0,0854 %**

Recall

TP/(TP+FN) = 22/(22+16) = **0,5789 %**

F-measure

2*(Precision*Recall)/(Precision+Recall) = 2*(0,6875*0,5789)/(0,6875+0,5789) = **1,2664**

➔ ROC curve and the area under the curve

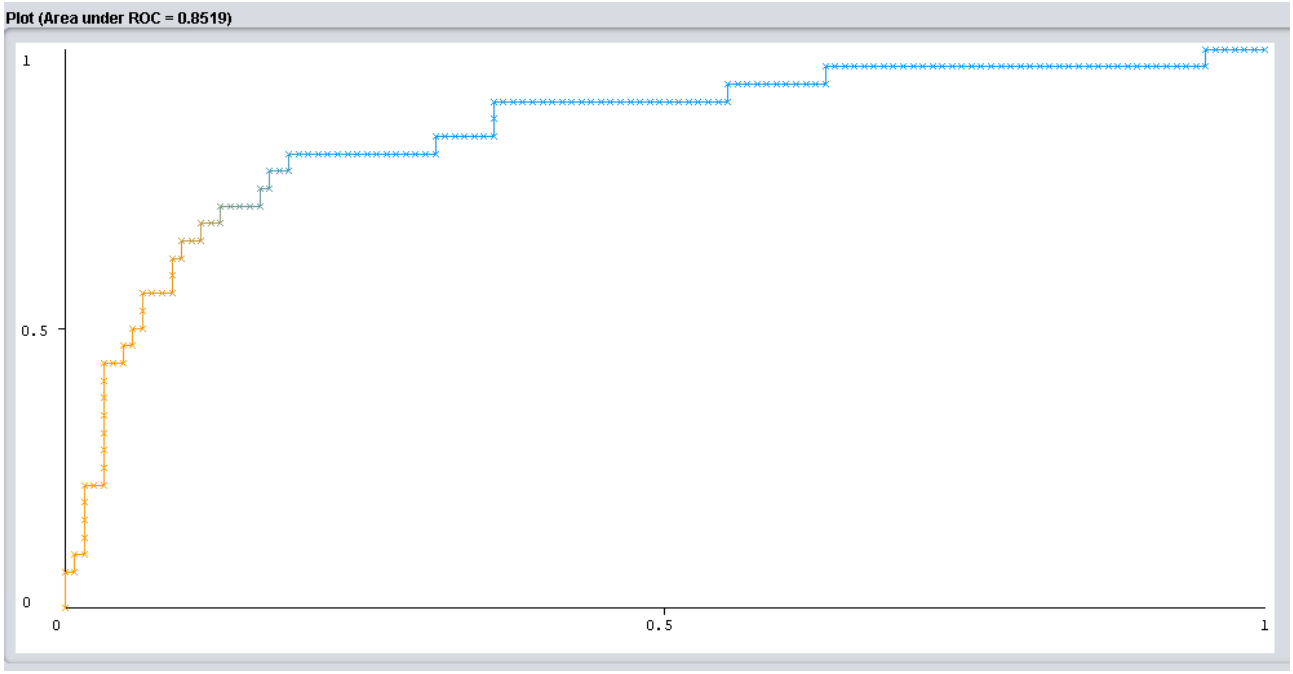

Ilustración 2: Graphic ROC Hepatitis

POST-OPERATIVE PATIENTS DATASET

➔ Introduction

The main idea of this classification problem is given a dataset about patients which have passed an operation and they are in a postoperative recovery area waiting to a decision of where they should be sent next. Depending on a number of temperature measurements we train a Naive Bayes classification model and evaluate its performance.

In this dataset we have 90 instances in total of which they are divided in three categories or classes, 2 of class I, 24 of class S and 64 of class A. Each instance has 9 attributes, one of them is the class attribute. Because of we have three possible classes we are faced with a multiclass classification problem.

➔ Number of classes

In this case our class attribute can take only **three *values***, I, S and *A*. Because of that we are faced to a **multiclass *classification problem***.

➔ Number of attributes

In this case we have **nine attributes** which *l-core, l-surf, l-02, l-bp, surf-stbl, core-stbl, bp-stbl, comfort* and *class.*

➔ Number of samples

In this case in particular the number of samples or instances is **90.**

➔ **Values of permance measures**

Confusion Matrix

=== Confusion Matrix ===

a	b	c	← classified as
62	0	2	a = A
2	0	0	b = I
23	0	1	c = S

Accuracy

Correctly Classified Instances	63
Incorrectly Classified Instances	27
Total instances	90

Acurracy = 63/90 = **70 %**

➔ Evaluation of the results

Dataset	Acurracy	Precision	Fallout	Recall	F-mesure	ROC
BreastCancer	72,7273	0,8507	0,4477	0,7808	1,6315	0,7034
Hepatitis	82,2258	0,6875	0,0854	0,5789	1,2664	0,8519
Post-Operative	70					

➔ Differences between datasets

Dataset	NumSamples	ClassType	ClassDistribution	Atributte Characteristics
BreastCancer	286	Binary	201/85	Linear/Nominal
Hepatitis	155	Binary	32/123	Nominal/ Numerical
Post-Operative	70	Multiclass	2/24/64	Nominal/ Numerical

➔ Last conclusions

If we look the performance results we can say that the Hepatitis dataset is the most suitable for the Naive Bayes classifier because this one has the best result in accuracy, area under ROC curve and fallout and althought in precision BreastCancer dataset presents a better result than him this could be due to the distribution of its classes, not being a realiable measure at all. In the case of PostOperatice dataset this has the lowest value of the three in the only performance measure that shows, accuracy, this bad result can be due to the disproportion between its classes or due to the small size of the dataset.